Supernova Recognition using Support Vector Machines

Raquel A. Romano Cecilia R. Aragon Chris Ding
Computational Research Division
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA 94720
E-mail: {romano,aragon,chqding}@hpcrd.lbl.gov

Abstract

We introduce a novel application of Support Vector Machines (SVMs) to the problem of identifying potential supernovae using photometric and geometric features computed from astronomical imagery. The challenges of this supervised learning application are significant: 1) noisy and corrupt imagery resulting in high levels of feature uncertainty, 2) features with heavy-tailed, peaked distributions, 3) extremely imbalanced and overlapping positive and negative data sets, and 4) the need to reach high positive classification rates, i.e. to find all potential supernovae, while reducing the burdensome workload of manually examining false positives. High accuracy is achieved via a sign-preserving, shifted log transform applied to features with peaked, heavy-tailed distributions. The imbalanced data problem is handled by oversampling positive examples, selectively sampling misclassified negative examples, and iteratively training multiple SVMs for improved supernova recognition on unseen test data. We present cross-validation results and demonstrate the impact on a large-scale supernova survey that currently uses the SVM decision value to rank-order 600,000 potential supernovae each night.

1. Introduction

One of the most important scientific discoveries of the last decade is that that the universe is expanding at an increasing rate [7, 8]. Type Ia supernovae play a critical role in this discovery because they provide measurements of the universe's expansion history since the time they began emitting light, several billion years ago. The scientific challenge of finding such supernovae stems from their rare occurrences (a few times each millenium in a typical galaxy), short lifespans (several weeks), and the necessity of detecting them as early as possible after explosion, before maximum brightness is reached. The computational challenge lies in the large volume of data that must be rapidly and accurately analyzed in order to find the 1 or 2 supernovae that could appear on a given night out of hundreds of thousands of potential candidates.

Current methods for finding supernovae do not employ machine learning techniques. Existing systems typically compute photometric and geometric features of subimages thought to contain potential supernovae, threshold each of these features, and send those subimages that satisfy all thresholds to human scanners who reject or accept them for follow-up studies. Due to high and variable image noise levels and image processing artifacts, the majority of subimages that pass the thresholds are false alarms that humans must manually reject. The thresholds are frequently adjusted in an attempt to minimize the number of subimages that must be viewed by humans while guaranteeing that no real supernovae are rejected. These manually tuned thresholds are brittle and often result in both missed supernovae and too many false positives for human scanners to evaluate daily.

There is an obvious need for supervised learning techniques to replace thresholding schemes for recognizing likely supernova candidates in digital sky surveys. Because feature distributions may change significantly over time due to the changing phase of the moon and frequent modifications to the data acquisition procedures (e.g., new telescopes, equipment changes, calibrations, weather changes, image processing improvements), there is a need for a systematic and principled way to adjust the decision criteria, which classifier training provides. Support Vector Machines are particularly appropriate because the high overlap between supernovae and non-supernovae in the computed feature space calls for a decision boundary that is nonlinear in the original space. Because ongoing surveys contain a vast and growing data set of labeled examples, supervised learning is a viable solution.

Statistical learning algorithms are still largely unused in the domain of astronomical object recognition from large digital sky surveys. Supervised learning techniques such

as SVMs and neural networks have been applied to galaxy and star classification [6] and AGN (active galactic nucleus) identification [12]. Previous work in applying SVMs to large-scale, imbalanced data in other domains are found in [2], who use replication to handle imbalanced data sets, and [11, 3, 4, 5], who use SVMs in combination with minority class subsampling, incremental training, boosting, and/or ensemble learning.

This work presents the first SVM application to the recognition of supernovae from astronomical imagery. We apply a novel feature transformation to handle peaked, heavy-tailed feature distributions, and demonstrate the necessity of this transform for obtaining high recognition rates. An incremental sampling and training strategy is adopted to handle the extreme imbalance in positive and negative samples (more than 10,000 negatives for every positive, due to the rare occurrences of Type Ia supernovae and the large sets of data gathered nightly). Positive data are oversampled by replication, and negative data are undersampled by using misclassified negatives to iteratively retrain the model and refine the decision boundary. We show that incremental sampling improves the trade-off between recognition rate and false positive rate, significantly reducing the daily human workload.

Section 2 describes the scientific mission of supernova search, the software that generates the image data in our study, and the features extracted from subimages to describe potential supernovae. Section 3 defines the feature transformation applied to peaked, heavy-tailed distributions. Section 4 describes the sampling strategies used to train and optimize the SVM. Section 5 describes empirical studies showing performance improvements gained by feature transformation and incremental sampling and training. Finally, Section 6 discusses the current impact of this work on a large-scale supernova survey that uses our SVM decision value to rank up to 600,000 potential supernovae per night.

2. Background

This section describes the scientific mission of the search for supernovae and outlines the existing methods used by astrophysicists to semi-automatically find potential supernovae in large image sets captured on a nightly basis.

The Nearby Supernova Factory (SNfactory) is an international project to obtain spectrophotometry data on a large sample of Type Ia supernovae in a "nearby" redshift range ($0.03 < z < 0.08$, or about 0.4 to 1.1 billion light years away) in order to measure the expansion history of the universe [1]. Such stellar explosions are very rare, occurring only a couple of times per millenium in a typical galaxy, and remaining bright enough to detect for only a few weeks. Previous studies of Type Ia supernovae led to the discovery of the mysterious "dark energy" that is causing the universe to expand at an accelerating rate.

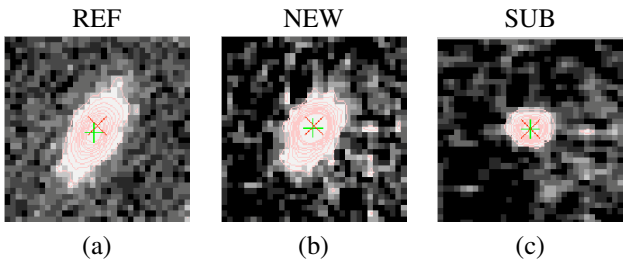

Figure 1: Example subimages from which geometric and photometric features are computed for SVM training and prediction. (a) Stacked reference images from previous nights, showing a galaxy; (b) Newly captured image (note different noise levels and image resolutions); (c) Subtraction of reference from new images, showing a supernova in a host galaxy.

To reduce the statistical uncertainties in previous experimental data, extensive spectral and photometric monitoring of more Type Ia supernovae is essential. The SNfactory collaboration has built an automated system consisting of specialized software and custom-built hardware that systematically searches the sky for new supernovae, screens potential candidates, then performs multiple spectral and photometric observations on each supernova. These observations will be stored in a database to be made available to researchers world-wide for further study and analysis.

Each night, the SNfactory receives about 80 GB of wide field CCD imaging data taken by the Quest-II camera on the Oschin 1.2-m telescope on Mt. Palomar for the Near Earth Asteroid Tracking (NEAT) project at JPL. Images are transferred from Mt. Palomar via the High Performance Wireless Research and Education Network (HPWREN) to the High Performance Storage System (HPSS) at the National Energy Research Scientific Computing Center (NERSC) in Oakland, California. Each morning, SNfactory search software running on NERSC's 700-node computing cluster, the Parallel Distributed Systems Facility (PDSF), matches images of the same area of the sky, processes them to remove noise and CCD artifacts, then performs an image subtraction from previously observed reference images on each set of matched images (Figure 1) [10].

A set of 19 features (Table 1) are computed on subimages of each subtraction image in which bright, point-source-like objects are detected. In the existing software, each feature is thresholded from above and/or below, and subimages that satisfy all thresholds are identified as containing a potential supernova candidate and saved to a database. Several hundred images containing one or more candidate subimages are saved each night.

Feature Name	Feature Definition
apsig	signal-to-noise ratio in aperture
perinc	% flux increase in aperture from REF to NEW
pcygsig	difference of flux in 2*FWHM of aperture and 0.7*FWHM; detects misaligned REF and NEW images)
mxy	x-y moment of candidate
fwx	FWHM of candidate in x
fwy	FWHM of candidate in y
neighbordist	distance to the nearest object in REF
new1sig	signal-to-noise of candidate in NEW1
new2sig	signal-to-noise of candidate in NEW2
sub1sig	signal-to-noise of candidate in SUB1
sub2sig	signal-to-noise of candidate in SUB2
sub2minsub1	weighted signal-to-noise difference between SUB1 and SUB2
dsub1sub2	difference in pixel coordinates between SUB1 and SUB2 (motion measurement)
holeinref	measure of negative pixels on REF in region of candidate
bigapratio	ratio of sum of positive pixels to sum of negative pixels within aperture
relfwx	REF image FWHM in x divided by NEW image FWHM in x
relfwy	REF image FWHM in y divided by NEW image FWHM in y
roundness	object contour eccentricity; ratio of powers in lowest order negative and positive Fourier contour descriptors
wiggliness	object contour irregularity; power in higher order Fourier contour descriptors divided by total power

Table 1: Definitions of features computed from image regions containing potential supernovae. Additional definitions: aperture = subarea of image within which features are computed; REF = reference image of a piece of sky composited from multiple nights; NEW1 (NEW2) = new images of same piece of sky; SUB1 (SUB2) = subtraction of NEW1 (NEW2) and REF images; FWHM = full-width half-max; flux = integrated photon levels over image subarea, i.e., sum of all pixel values within aperture.

These subimages are then visually scanned by humans, a process that takes approximately 8 person-hours for each night's data. Of these visually scanned subimages, about 1% are flagged as containing potential supernova candidates. These candidates are sent to the Supernova Integral Field Spectrograph (SNIFS) on the University of Hawaii 2.2-m telescope on Mauna Kea for spectrophotometric screening and follow-up. The resulting image and spectral data are stored in a database for scientific study.

The bottleneck in this process is the large amount of manual labor required to find good supernova candidates. Feature thresholds are often adjusted manually (known as "tightening the cuts") in order to reduce the number of bad image subtractions sent to human scanners. However, this results in the occasional loss of a good supernova candidate, and due to their rarity, it is important to miss as few candidates as possible.

3. Feature Transformation

Of the 19 features extracted from subimages (Table 1), 10 features (apsig, perinc, pcygsig, mxy, neighbordist, new1sig, new2sig, sub1sig, sub2sig, holeinref) were found upon inspection of sample histograms, to have peaked, heavy-tailed, and often skewed distributions. A simple log transform, often used to transform kurtotic distributions to more Gaussian-like distributions, could not be used since the supernova features may take on negative values. Instead we apply a modified log-transform that preserves the feature's sign, and is well-defined for all real values:

$$f(x) = \mathrm{sgn}(x)\log(1+|x|)$$

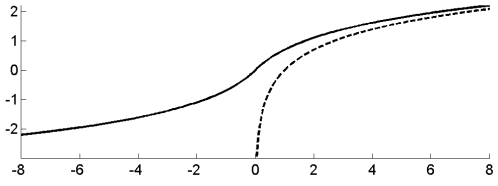

Figure 2: Sign-preserving, shifted log transform for normalizing features with peaked, heavy-detailed distributions. Solid line: $f(x) = \mathrm{sgn}(x)\log(1+|x|)$; dashed line: $f(x) = \log(x)$.

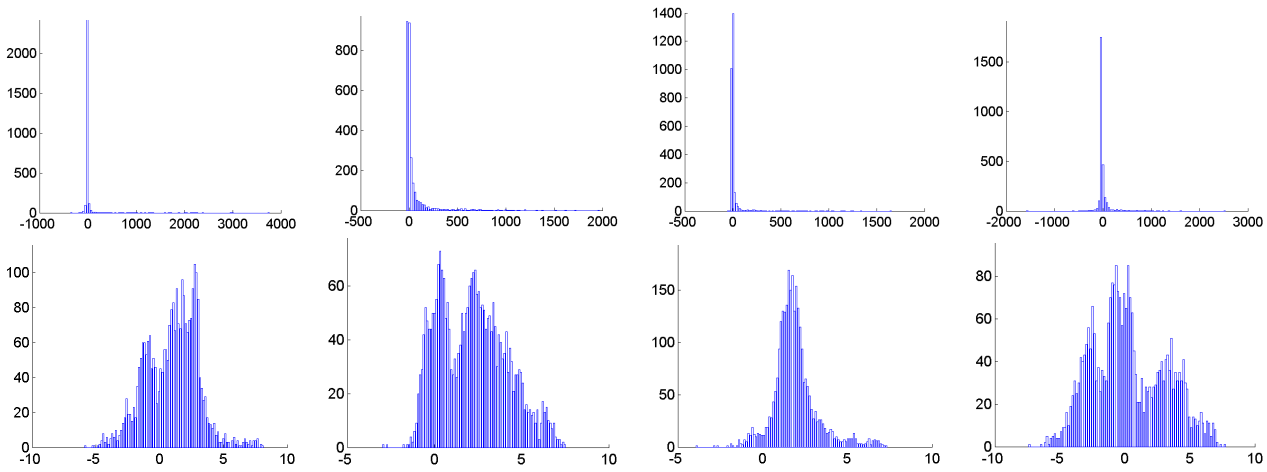

Figure 3: Transformation of 4 different features: (a) `pcygsig`, (b) `new1sig`, (c) `sub1sig`, and (d) `holeinref`. For each feature, the top histogram shows the original distribution of raw data, and the bottom histogram shows the distribution of the transformed data, as described in Section 3. The skewed, sharply peaked, heavy-tailed distributions are transformed to Gaussian-like distributions that are more amenable to statistical analysis.

Figure 2 illustrates the shape of the feature transform function, and Figure 3 shows sample histograms of features with heavy-tailed, peaked distributions and their histograms after transformation. Application of the transform improves the positive and negative classification rates while preventing overfitting of the data, as shown in Section 5.1.

4. Supernova Recognition

The SVM algorithm is a classification method that has successfully been applied to many pattern recognition problems and is founded on an elegant mathematical theory of statistical learning [9]. It finds an optimal hyperplane parameterized by a normal vector $\mathbf{w}$ and offset b in a high-dimensional feature space that separates data samples belonging to two different classes. The hyperplane defines the decision boundary maximizing the margin between data samples in the two classes, therefore giving good generalization to new data samples. The decision boundary is defined in terms of the hyperplane as the function $f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b$, where ϕ is a function that embeds the problem into a higher-dimensional feature space in which the classes are more easily separable than in the original feature space. A sample data point $\mathbf{x}$ is typically classified into one of the two classes by thresholding $f(\mathbf{x})$. The actual embedding is achieved through a kernel function defining an inner product in the embedding space, $k(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1) \cdot \phi(\mathbf{x}_2)$, which makes class prediction easy to implement and fast to compute.

Due to a highly imbalanced data set, the choice of sample set for SVM training is nontrivial, and performance on unseen data is highly sensitive to the samples used in building the model. A typical night yields 300,000-600,000 new examples, among which 5-30 (.001%-0.01%) are positive examples. Of these positive examples, only several are actual supernovae and the rest are other transient objects that appear in the image subtractions, e.g. asteroids, variable stars, and AGNs, or imaging artifacts that resemble transient objects, both in the feature space and when visually scanned.

Two sampling strategies are used to address the imbalanced data problem. First, oversampling of positives by replication with random undersampling of negatives prevents the SVM from favoring negative performance. As shown in Section 5, positive oversampling by a factor of 10 and negative undersampling by a factor of about .001 to construct a balanced training set, result in cross-validation rates of 97% positive classification and 96% negative classification.

To improve the recognition of supernovae in new data, we use an incremental sampling strategy to iteratively train a new model using negative samples that are misclassified with high decision values by a previously trained model. Such boosting-like training schemes for SVMs have been shown to be successful by [3, 4, 5]. We train an initial model on a balanced set of randomly sampled negatives and 10-fold oversampled positives. This model is then tested on a new set of negative examples, which are ranked by descending SVM decision value. The top-ranked negatives are chosen for a new training set, along with enough randomly sampled negatives to match the oversampled positive set. Section 5.2 demonstrates the SVM refinement's ability

to decrease the false positive rate to less than 1%.

5. SVM Optimization

Optimal SVM parameter selection is accomplished by 5-fold cross-validation and grid search. Three types of SVMs and kernels were initially tested (C-SVM+RBF kernel, ν-SVM+RBF kernel, C-SVM+degree 3 polynomial kernel), giving similar cross-validation results; all remaining tests use a C-SVM with RBF kernel. The training set contains 200 positive examples of supernovae, variable stars, and any other celestial objects saved by human scanners during a 2-week period in May. In the same time period, there are several million negative examples (several hundred thousand examples per night of observation) from which to sample. Positives are oversampled by a factor of 10 to yield 2000 positives, and 2000 negatives are randomly subsampled by a factor of roughly 0.1% to create a balanced training set of 4000 examples. Table 2 shows 5-fold cross-validation rates using several different parameter settings.

C-SVM+RBF 5-Fold Cross-Validation			
Parameters	"+"	"-"	SV
$C = 0.5, \gamma = 2.0$	97% (1.2%)	96.3% (.8%)	15%
$C = 8.0, \gamma = 0.13$	97% (2.1%)	94% (1.3%)	16%
$C = 8.0, \gamma = 2.0$	92% (6.7%)	98% (.6%)	6%
$C = 128\ \gamma = 0.13$	95% (2.8%)	96.7% (.7%)	9%

Table 2: Positive and negative cross-validation rates using parameters chosen by grid search for C and γ. Standard deviations from the mean classification rates are given in parentheses. Percentages of training examples used as support vectors are shown in the last column.

5.1. Feature Transformation Experiments

The effectiveness of the feature transformation described in Section 3 is tested using the parameter grid search, data sampling, and cross-validation described above. Without application of the shifted log transform, cross-validation rates for the best parameters settings dropped by an average of 5% for positive classification and an average of 14% for negative classification. At the same time, the average percentage of training examples used as support vectors rose by an average of 23%, suggesting that the model suffers from overfitting when trained on untransformed data. A more informative comparison examines the trade-off between positive and negative classification rates as the decision value varies. Figure 4 shows improved ROCs (receiver operating curves) for several parameter settings when training is performed on transformed features as opposed to raw feature values.

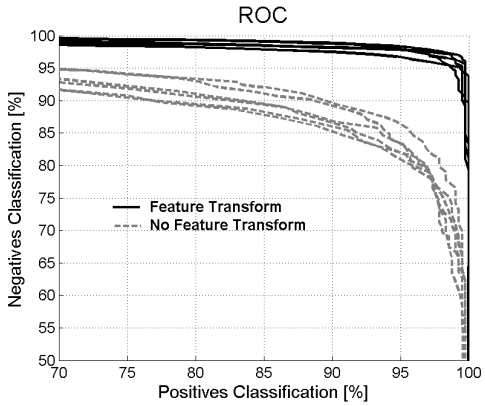

Figure 4: Receiver operating curves for several parameter settings. Models trained on transformed feature values (dark solid lines) out-perform models trained on raw feature values (gray dashed lines).

5.2. Incremental Sampling

In practice, the cross-validation performance reported above must be improved upon due to the large number of new negative examples generated each night. If a typical night of observation returns 500,000 negative examples, a 98% classification rate returns 10,000 false positives, far too many for human scanners to view each day. Even a 1% increase in negative classification rate to 99% would reduce the number of false positives by several thousand.

We leverage the availability of a large negative data set for reducing the number of false positives. Starting with the training set described in the previous section, we train a base model with parameters $C = 0.5$ and $\gamma = 2$, and validate its performance on an unseen test set of 417 positive examples found in June and July, and 5000 randomly sampled negatives, resulting in 99.5% positive and 96.4% negative classification. The base model is then applied to a new, training set of 5000 negative examples, and these negatives are ranked by descending SVM decision value. The original training set is then modified by removing half of the randomly sampled negatives and replacing them with the 1000 misclassified negatives with highest decision values from the new training set. A new model is trained and applied to another new training set to obtain a new set of misclassified negatives with high decision values. This procedure is iterated several times, each time choosing highly-ranked, misclassified negatives from a new set of negative training examples. The resulting series of models is tested on the June/July test set, and Figure 5 shows the improvement in negative performance rate by about 1%, the equivalent of several thousand fewer false positives on a typical night, at a corresponding positive recognition rate of 92.3%. Further understanding of the types of positive errors is enabled by

the fact that all positives are classified during visual scanning into categories indicating whether they were verified by humans to be actual supernovae as opposed to other celestial objects or so-called junk. Of the 22 positive errors, only one example was an actual supernovae, the rest being asteroids, variable stars, AGNs, or of unknown type.

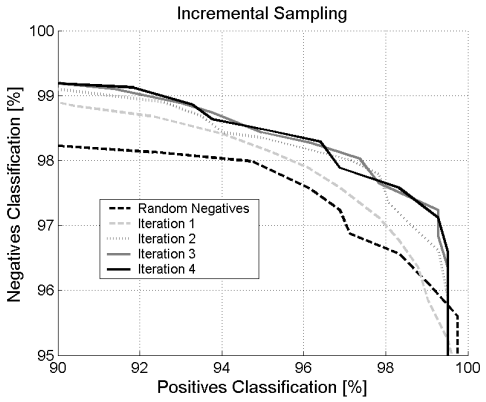

Figure 5: Improvement by incremental sampling. Initial model trained on randomly sampled negatives (dashed dark line) is improved upon by iteratively retraining with misclassified negatives with high decision values. A 1% improvement in negative classification rate (solid dark line) reduces the number of false positives by several thousand on a typical night.

6. Discussion

This work describes a novel application of Support Vector Machines to a problem in astrophysics that is in growing need of statistical learning methods: the search of increasingly large collections of digital imagery for astronomical objects. The difficulties brought about by noisy imagery, imbalanced data, overlapping classes, and unusual feature distributions raise interesting technical challenges that preclude straightforward application of SVMs. We demonstrate the importance of applying a sign-preserving, shifted log transform to certain features, and the necessity of iteratively training and selectively sampling the data in order to refine the decision boundary in the face of unbalanced classes. Frequent changes in data acquisition procedures exacerbate the difficulty, since the set of positive examples disappears every time operations change, so the need to quickly bootstrap a new model is essential. Comparisons between SVMs and ensemble learners such as boosted decision trees and random forests are currently underway.

The SVM has been integrated into a large-scale supernova survey that currently receives hundreds of thousands of subimages per night, of which only a handful are expected to contain real supernovae. Astrophysicists who in the past have visually scanned thresholded candidates to eliminate false positives now rank-order all candidates by SVM decision value so that likely supernovae are immediately viewed, and the effective scanning load has been reduced to less than 1 person-hour per day. This capability allows scientists to quickly discover new supernova candidates while drastically reducing the burden of visual scanning. The resulting efficiency improvements demonstrate the great impact that supervised learning may have on future digital sky surveys that are slated to collect orders of magnitude more imagery in search of far more numerous and faint celestial objects.

References

[1] G. Aldering, G. Adam, P. Antilogus, et al. Overview of the Nearby Supernova Factory. In J. A. Tyson and S. Wolff, editors, *Survey and Other Telescope Technologies and Discoveries*, volume 4836 of *Proceedings of the SPIE*, pages 61–72, December 2002.

[2] C. Ding and I. Dubchak. Multi-class Protein Fold Recognition Using Support Vector Machines and Neural Networks. *Bioinformatics*, 17:349–358, April 2001.

[3] Z. Erdem, R. Polikar, F. S. Gürgen, and N. Yumusak. Ensemble of SVMs for Incremental Learning. In *Multiple Classifier Systems*, pages 246–256, 2005.

[4] P. Kang and S. Cho. EUS SVMs: Ensemble of Under-Sampled SVMs for Data Imbalance Problems. In *Lecture Notes in Computer Science Series*, Proceedings of the 13th International Conference on Neural Information Procerssing (ICONIP), October 2006.

[5] Y. Liu, A. An, and X. Huang. Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. In *Lecture Notes in Computer Science*, volume 3918 of *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, April 2006.

[6] Z.-Y. Liu, K.-C. Chiu, and L. Xu. Improved System for Object Detection and Star/Galaxy Classification via Local Subspace Analysis. *Neural Networks*, 16(3-4):437–451, 2003.

[7] S. Perlmutter, G. Aldering, G. Goldhaber, et al. Measurements of Omega and Lambda from 42 High-Redshift Supernovae. *Astrophysics Journal*, 517:565–586, 1999.

[8] A. G. Riess, A. V. Filippenko, et al. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *Astronomical Journal*, pages 1009–1038, 1998.

[9] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[10] W. M. Wood-Vasey. *Rates and Progenitors of Type Ia Supernovae*. PhD dissertation, U. of California at Berkeley, 2004.

[11] H. Yu. SVM Selective Sampling for Ranking with Application to Data Retrieval. In *KDD '05: 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 354–363, New York, NY, USA, 2005. ACM Press.

[12] Y. Zhang and Y. Zhao. Automated Clustering Algorithms for Classification of Astronomical Objects. *Astronomy and Astrophysics*, 422:1113–1121, Aug. 2004.